Binary Regression

STAT 245

Data Source

The dataset used here is on Alaskan wood frogs, detailing some physical characteristics, habitat characteristics, and the number of developmental and other abnormalities found in the frogs. It was originally obtained from:

http://datadryad.org/resource/doi:10.5061/dryad.sq72d.

Data Source

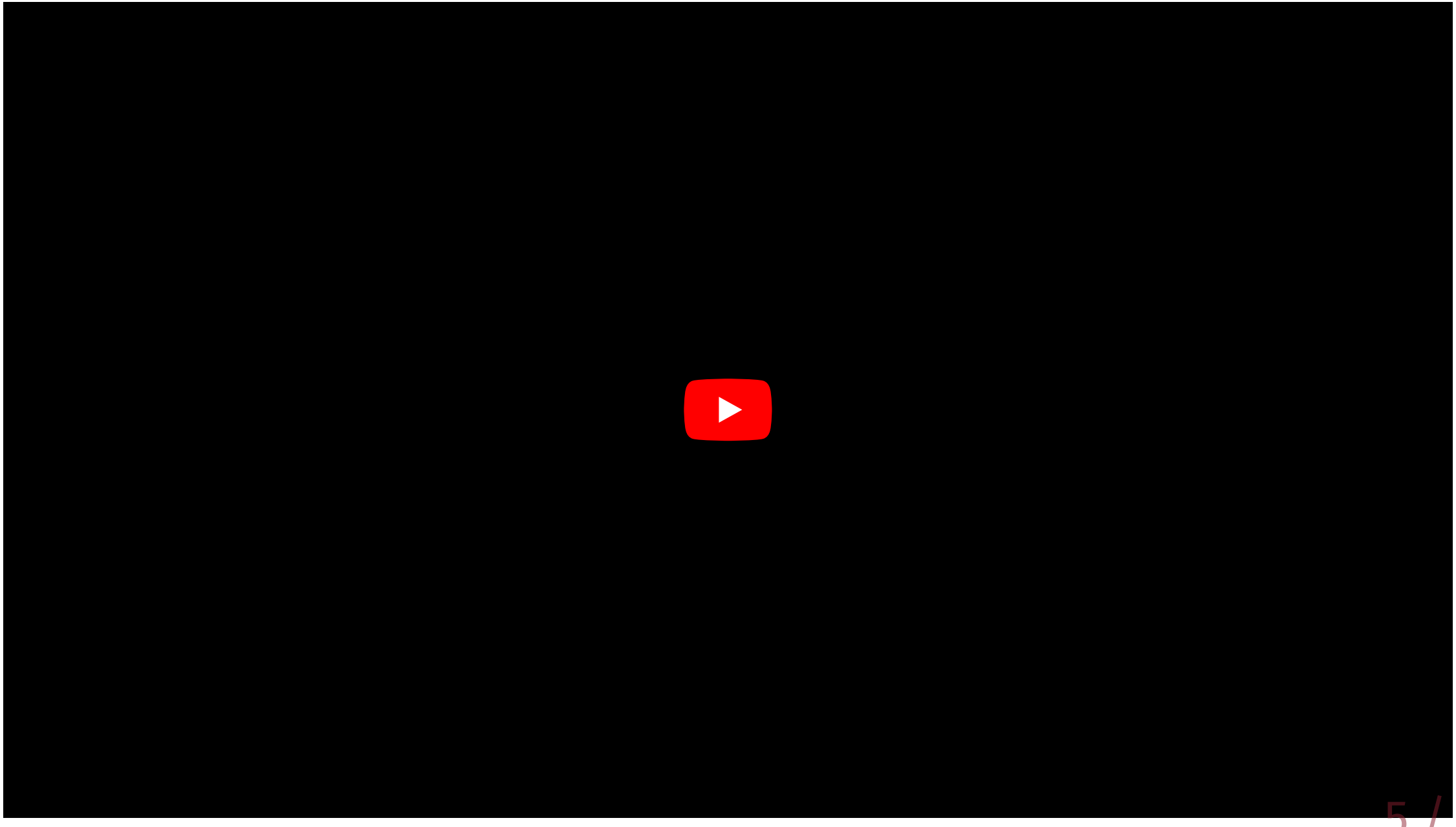
The data file can be accessed online at:

https://sldr.netlify.app/data/FrogAbnormalities.csv

```
frogs <- read_csv(
  'http://sldr.netlify.com/data/frog-abnormalities.csv')
DT::datatable(frogs, width = 500)
```

Show 10 entries Search:

	collection_id	frog_id	gosner_stage	tail_length	frog_comments	abnormal	bleeding_injury	skeletal_abnormality
1	KNA1021-RASY-080712	15	stage 45	2		No	No	No
2	KNA1024-RASY-080812	13	stage 44	20		No	No	No
3	KNA1069-RASY-080612	24	stage 45	1		No	No	No
4	KNA1090-RASY-080612	5	stage 45	3		No	No	No
5	KNA11119-RASY-081612	26	stage 44	17		No	No	No
6	KNA1024-RASY-080812	47	stage 44	33	~ 3mm of right thigh is comparable to left, remainder of thigh/calf are underdeveloped and foot is not fully developed,	Yes	No	Yes



Variables in the dataset include:

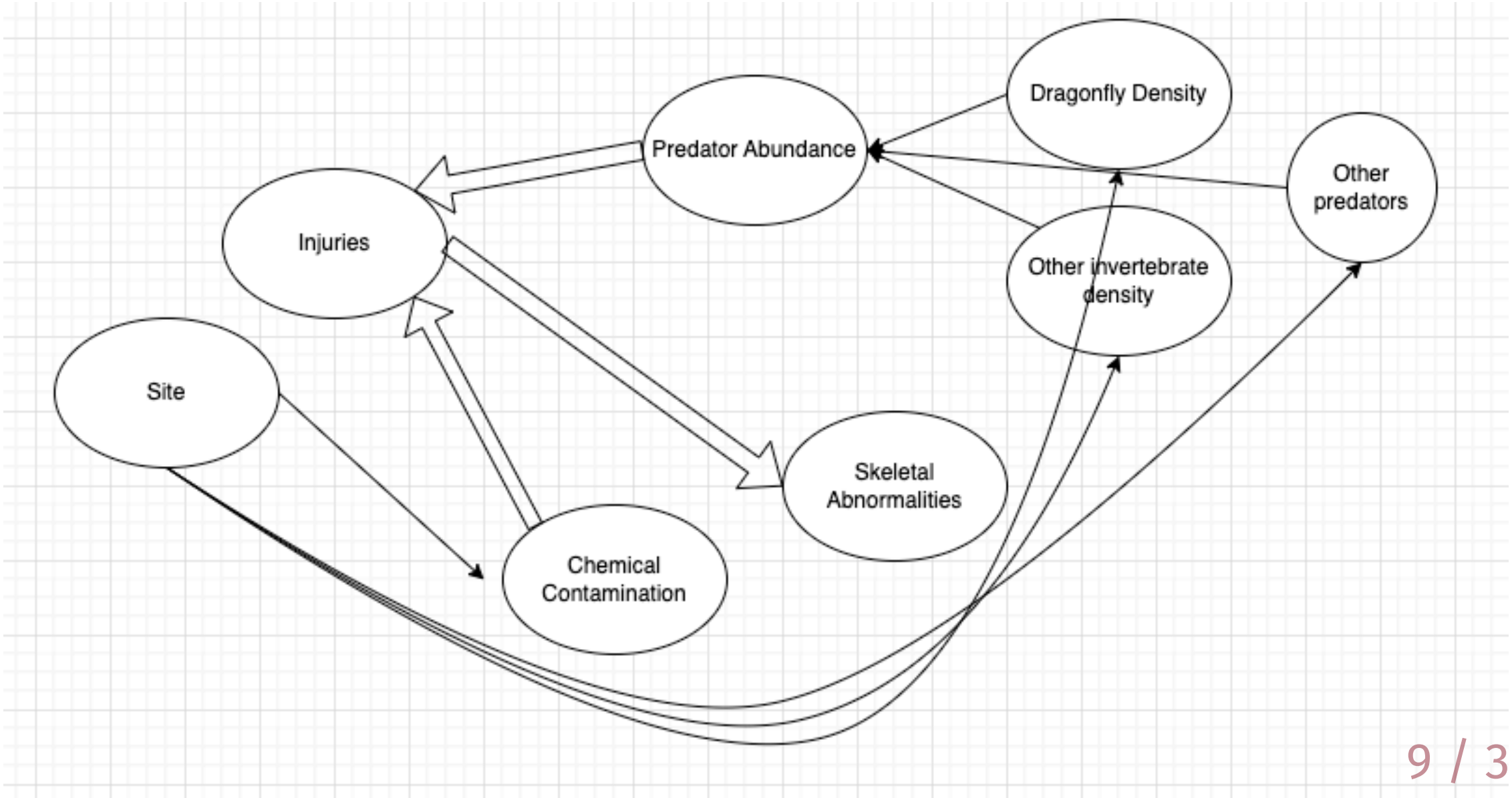
- IDs: `collection_id`, `frog_id`, `site`
- Info on time of data collection: `date`, `year`, `coll_date`
- Size and developmental stage of the frog: `gosner_stage`, `tail_length` (which is longer for young frogs, that is, tadpoles)
- Whether or not the frog has any abnormality in general (`abnormal`), an injury (`bleeding_injury`), or a specific type of abnormality: `skeletal_abnormality`, `eye_abnormality`, `surface_abnormality`
- Relative abundance of invertebrate predators of frogs: `dragonfly_relative_density` and `other_invert_relative_density`
- *Rough* water testing results: `detectable_analytes` (average number of contaminants present)

Dataset Size: n/15 revised

Model Plan?

The repeated occurrence of abnormal amphibians in nature points to ecological imbalance, yet identifying causes of these abnormalities has proved complex. Multiple studies have linked amphibian abnormalities to chemically contaminated areas, but inference about causal mechanisms is lacking. Here we use a high incidence of abnormalities in Alaskan wood frogs to strengthen inference about the mechanism for these abnormalities. We suggest that limb abnormalities are caused by a combination of multiple stressors. Specifically, toxicants lead to increased predation, resulting in more injuries to developing limbs and subsequent developmental malformations. We evaluated a variety of putative causes of frog abnormalities at 21 wetlands on the Kenai National Wildlife Refuge, south-central Alaska, USA, between 2004 and 2006. Variables investigated were organic and inorganic contaminants, parasite infection, abundance of predatory invertebrates, UVB, and temperature. Logistic regression and model comparison using the Akaike information criterion (AIC) identified dragonflies and both organic and inorganic contaminants as predictors of the frequency of skeletal abnormalities. We suggest that both predators and contaminants alter ecosystem dynamics to increase the frequency of amphibian abnormalities in contaminated habitat. Future experiments should test the causal mechanisms by which toxicants and predators may interact to cause amphibian limb abnormalities. - Reeves et al. 2010, https://doi.org/10.1890/09-0879.1

Causal Diagram?



Critique?

```
skeletal_abnormality ~ detectable_analytes +
  dragonfly_relative_density +
  other_invert_relative_density
```

- Why not include injuries?
- Would you include site?
- Why isn't the frog's developmental stage in there?

Regression Evolution

old: linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots \beta_k x_k + \epsilon$$

- where xs are the k predictor variables,
- βs are the parameters to be estimated by the model,
- and $\epsilon \sim N(0, \sigma)$ are the model residuals.

Regression Evolution

When our response variable was a *count* variable, we modified our equation to:

$$log(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 +$$

$$\ldots \beta_k x_k + \epsilon_{link}$$

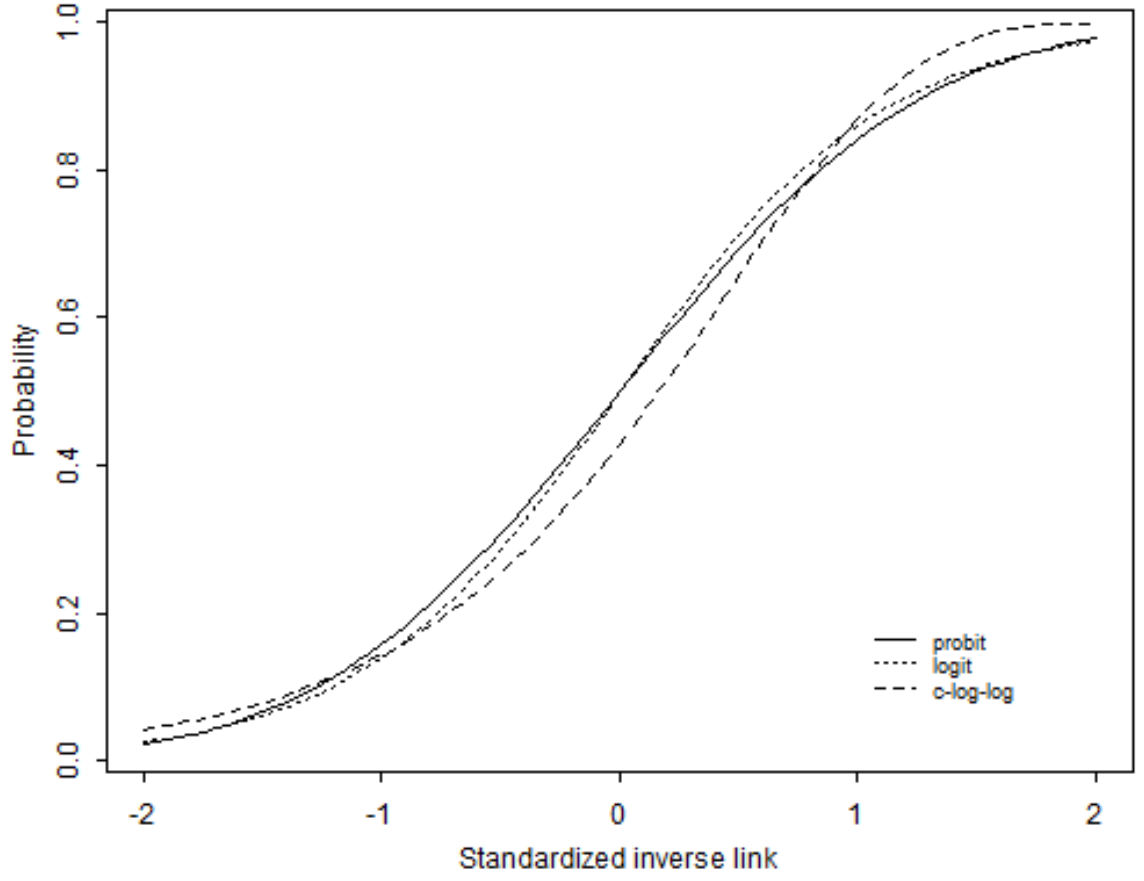
positing that $y_i \sim Pois(\lambda_i)$ for Poisson regression; similarly for negative binomial regression, we just replaced that Poisson distribution with a negative binomial distribution (and replaced λ_i with μ_i, stating that $y_i \sim NegBin(\mu_i, \sigma)$).

Binary Response?

What if our response variable is *logical* -- a categorical variable with just two possible values? We will designate one of the two values a "success," and then we want to predict the probability of success as a function of some set of predictors. What will our model equation look like in this case?

Which Distribution?

Link Functions



Note: figure is from http://data.princeton.edu/wws509/notes

Binary Regression Equation

Back to Frogs...

How does this equation relate back to our desired response variable?

- y_i, the ith observation of the response variable is assumed to follow a binomial distribution with probability p_i
- In other words: $y_i \sim Binom(n_i, p_i)$
- n_i depends on the setup of the data -- often n = 1 for each row of the dataset, as here where each row is one frog. We can think of each frog as one binomial trial, with success/failure meaning abnormality/normality of the frog.

Checking the data setup

- We would like to model the proportion frogs with abnormalities as a function of a set of predictors.
- The variable `skeletal_abnormality` has values "Yes" and "No".
- In R, if we use this (categorical) variable as our response, how will R determine which level (value of the variable) is a "success"?

Response Variable

```
frogs |>
  # pull out just the variable in question
  pull(skeletal_abnormality) |>
  # print out the variable's unique values
  levels()
```

```
## NULL
```

`factor()` Response Variable

```
frogs |>
  # force our response to be "factor" not "character".
  # This will also auto-sort the levels in to alpha order.
  # So don't do it ever AFTER you have carefully re-ordered  them!
  mutate(skeletal_abnormality = factor(skeletal_abnormality)) |>
  # pull out just the variable in question
  pull(skeletal_abnormality) |>
  # print out the variable's unique values
  levels()
```

```
## [1] "No"  "Yes"
```

Better: Rearrange Levels?

```
# list the values in the "new" order you want
# (here it won't actually change from the original...
# but this is how you *would* change it if needed)
frogs <- frogs |>
  mutate(skeletal_abnormality = forcats::fct_relevel(skeletal_abnormality, 'No', 'Yes'))
# check again
frogs |> pull(skeletal_abnormality) |> levels()
```

```
## [1] "No"  "Yes"
```

Model Fitting

```
frog_model <- glmmTMB(skeletal_abnormality ~
                          detectable_analytes +
                          dragonfly_relative_density +

other_invert_relative_density,
                  data = frogs,
                  family = binomial(link = 'logit'))
```

```
summary(frog_model)
```

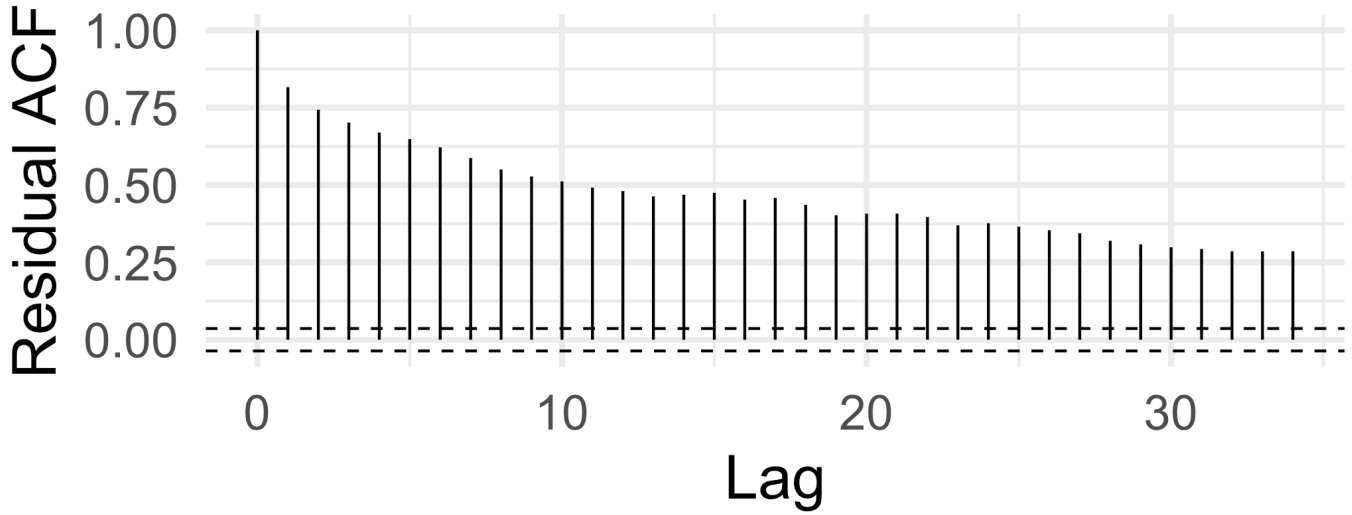
```
##  Family: binomial  ( logit )
## Formula:
## skeletal_abnormality ~ detectable_analytes + dragonfly_relative_density +
##     other_invert_relative_density
## Data: frogs
##
##      AIC      BIC   logLik deviance df.resid
##   1382.8   1406.8   -687.4   1374.8     2918
##
##
## Conditional model:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.802e+00  1.599e-01 -17.515  < 2e-16 ***
## detectable_analytes          -3.270e-03  6.075e-03  -0.538  0.59034
## dragonfly_relative_density    5.209e-03  1.949e-03   2.672  0.00753 **
## other_invert_relative_density -3.357e-05  1.073e-04  -0.313  0.75425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assessment

Conditions

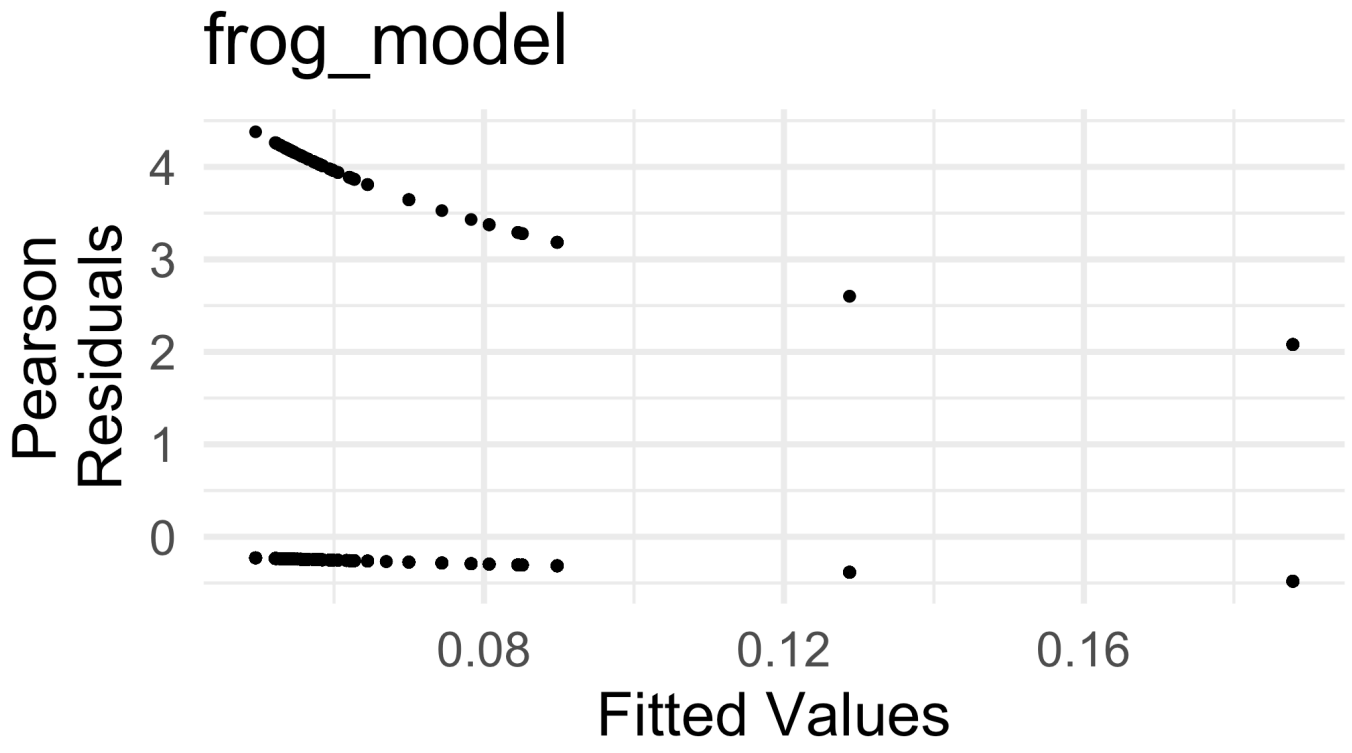
Assessment

ACF



Never Do This!

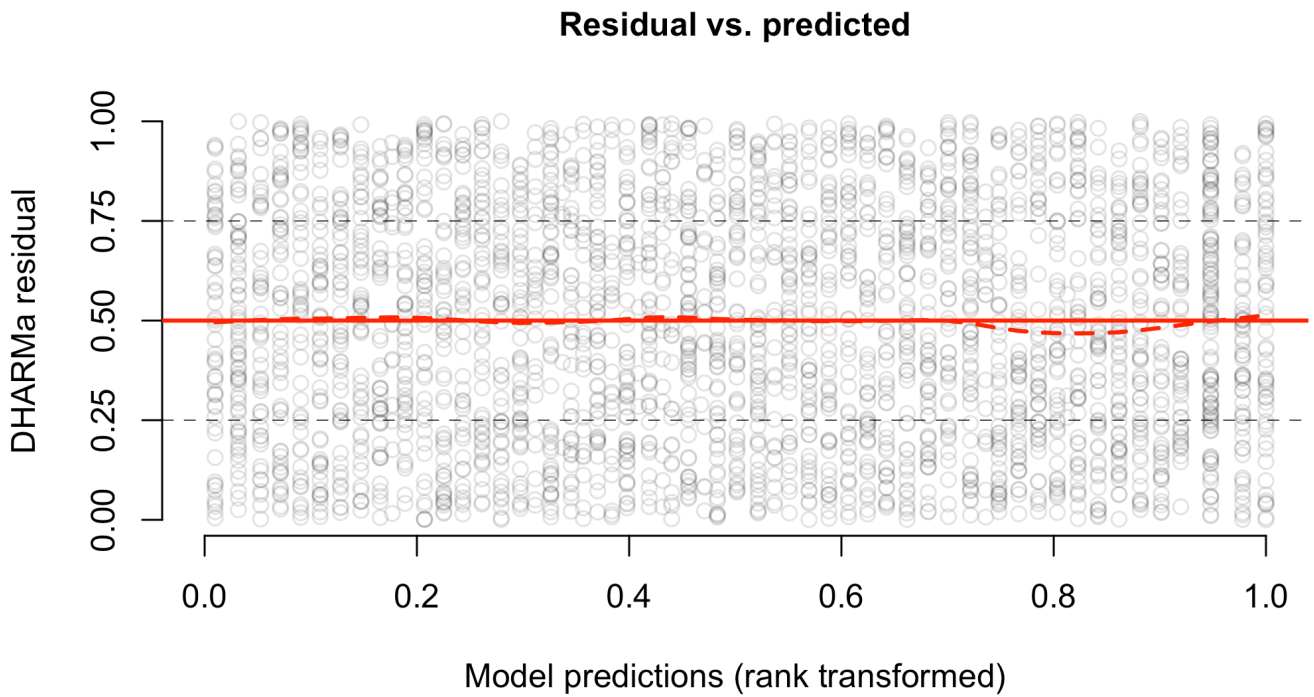
```
# DO NOT USE THIS IN YOUR OWN ASSESSMENT OF YOUR MODELS!
gf_point(resid(frog_model, type='pearson') ~ fitted(frog_model)) |>
  gf_labs(title='frog_model',
          y=' Pearson\nResiduals',x='Fitted Values')
```



Assessment

Scaled Residual Plot

```
# ACUTALLY USE **THIS** WAY FOR MEAN-VARIANCE CONDITION
library(DHARMa)
sim_frog_res <- simulateResiduals(frog_model)
plotResiduals(sim_frog_res, quantreg = FALSE)
```



Selection

Just as usual.

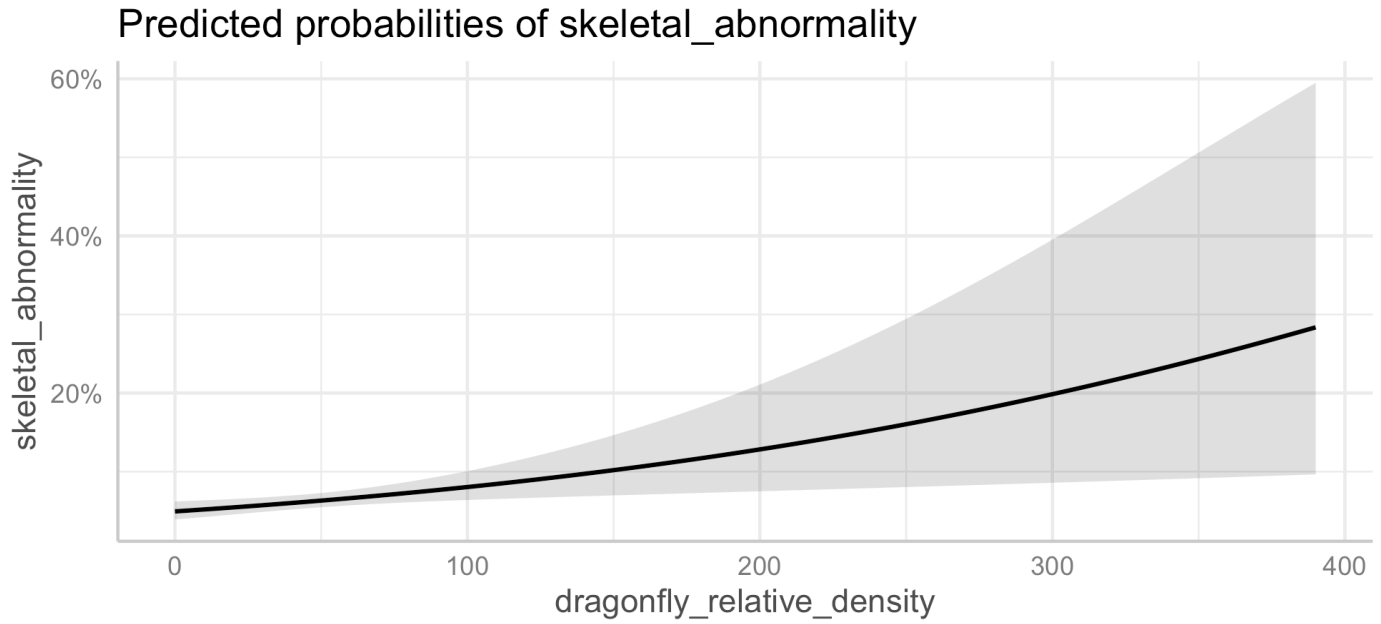
For example, using ANOVA:

```
car::Anova(frog_model)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: skeletal_abnormality
##                              Chisq Df Pr(>Chisq)
## detectable_analytes          0.2898  1   0.590339
## dragonfly_relative_density   7.1417  1   0.007531 **
## other_invert_relative_density 0.0980  1   0.754254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction Plots (same as ever)

```
library(ggeffects)
ggpredict(frog_model,
          terms = 'dragonfly_relative_density') |>
  plot()
```



Prediction Plots

```
ggpredict(frog_model,
          terms = 'detectable_analytes') |>
  plot()
```